Resource Related Research - Computers and Chemistry

ANNUAL REPORT
August 1, 1977 - April 30, 1978

Stanford University
NIH/BRP Grant RR-00612

Carl Djerassi, Principal Investigator
(Social Security No. [redacted]

1 OVERVIEW OF RESEARCH ACTIVITIES

In this first year of a three year renewal, substantial progress was made on every major item in the renewal proposal. The most obvious facets of this interdisciplinary work on computers and chemistry are research, engineering and applications. On the research side, the computer programs have grown in both chemical and computer science sophistication. On the engineering side, the programs have been made faster and easier to use. On the applications side, the programs have been used by chemists working on biomedical problems at Stanford and elsewhere as aids in their own research (see [4]). In this report we stress progress along the dimension of research, but mention the other aspects in the discussions of research progress.

The report is organized by the following problem areas:

Structure Elucidation
Theory Formation
C^{13}-NMR Problems
Collaborative Research
Instrumentation

Unpublished work is discussed in some detail, while published work is summarized here. The project continues at a vigorous pace and remains an exciting research atmosphere because of the unique collection of researchers dedicated to the goal of producing intelligent computer aids for biomedical research.

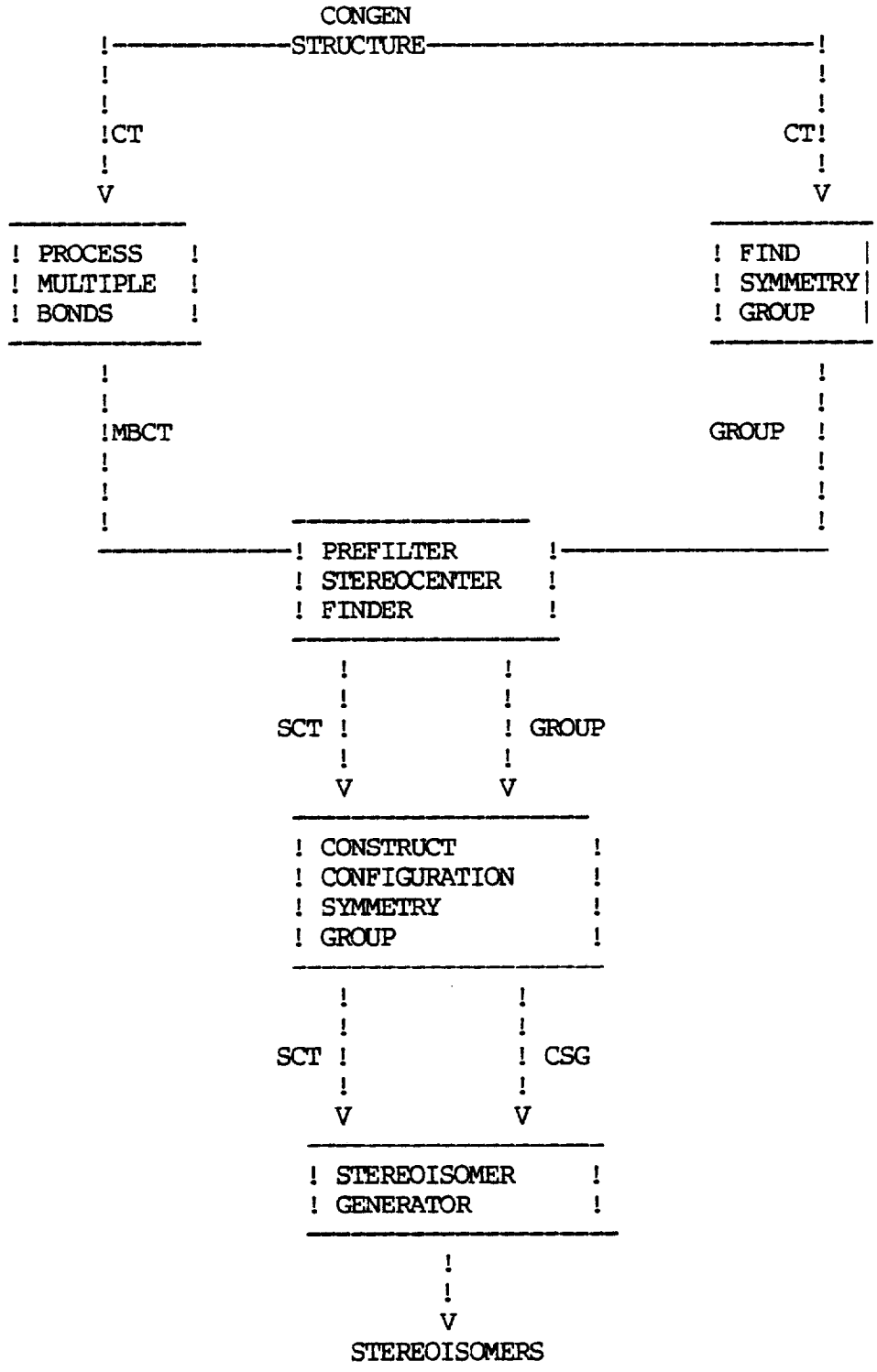
2 STRUCTURE ELUCIDATION PROGRAMS

2.1 Stereochemistry in CONGEN

The effort to give CONGEN the ability to recognize and use the stereochemical features of molecules in structure determination has continued for the past year. The proposed first stage in this effort was to write a program which was capable of recognizing the configurational stereochemical features of a molecule and generate all the possible stereoisomers based on these features. This program has been written and interfaced to an experimental version of CONGEN, and is described in detail below. The proposed second stage in this effort is to modify this program to permit generation of stereoisomers which satisfy certain constraints, much as the existing CONGEN program constrains the generation of topological isomers. This ongoing effort is discussed in the section on future plans.

Each module of this program, written in SAIL, is described in detail below. In summary, the program takes a structure defined in CONGEN and extracts the Connection Table (CT) from it. The symmetry group of this structure is found based on this connection table. The CT is then searched for features corresponding to multiple bond stereo features (double bonds, allenes, etc.) and the CT is modified to the Multiple Bond Connection Table (MBCT). Making use of the symmetry group, the MBCT is then searched for stereocenters (asymmetrically substituted carbon atoms, etc.) to yield the Stereochemical Connection Table (SCT). Using the SCT, the symmetry group is modified to recognize the effect of the symmetry operations on these stereocenters. The resulting group is the Configuration Symmetry Group (CSG). The SCT and the CSG are then used together to generate the possible stereoisomers for the input structure. These are output with other information in the manner described below.

Stereoisomer Generator Program


CONGEN
STRUCTURE
CT
CT
PROCESS MULTIPLE BONDS
FIND SYMMETRY GROUP
MBCT
GROUP
PREFILTER STEREOCENTER FINDER
SCT
GROUP
CONSTRUCT CONFIGURATION SYMMETRY GROUP
SCT
CSG
STEREOISOMER GENERATOR
STEREOISOMERS

2.1.1 Process Multiple Bonds

This module takes the CT and converts it into a Connection Matrix (CM) for use here and in the group finder described below. The CM is searched for all double and triple bonds. The atoms involved in triple bonds are flagged as stereochemically uninteresting. Double bonds and cumulenes with CH2 ends are similarly flagged. All remaining doubly-bonded atoms are potential stereocenters at this stage. These are processed by attaching a fictional bivalent node to each edge of the double bond, thus giving the multiply-bonded atom four distinct neighbors which aids in configuration assignment and in representation of the permutation group. These fictional nodes are given numbers higher than those already used in the structure and the corresponding rows are added to the connection table, to yield the Multiple Bond Connection Table (MBCT). (See examples.)

2.1.2 Find Symmetry Group

This module finds the node symmetry group of the input CT and was constructed largely of existing code from other parts of CONGEN, thereby saving the time and effort of developing another large program. This segment can be used independently from the rest of the program, a useful feature since previous group finders were written for very specific purposes. The symmetry group is constructed in two parts. The first is the node symmetry group of the input CT. The second is the symmetry group associated with the fictional nodes which were added to the MBCT described above. These two groups combine as a semidirect product. However, the utilization is such that the product group never needs to be explicitly constructed. This means the group can be stored in two arrays of size nXp and fXq where n is the number of original nodes, p is the order of the node symmetry group, f is the number of fictional nodes and q is the order of their symmetry group. If the entire group were constructed, the storage array would be of size nfXpq. Since the symmetry group can be by far the largest data structure in the program, the saving of space by this technique is crucial.

2.1.3 Prefilter

This module is all new code which recognizes all the stereochemically interesting features of the input structure based on the configuration of tetravalent atoms. The program works backwards by rejecting all those atoms which can never exhibit configurational stereochemistry. The MBCT is scanned first to eliminate all methyls and methylenes from further consideration as stereocenters. These atoms are flagged as nonstereocenters. Following this, all atoms with symmetrically related substituents are found using the node symmetry group described above. A crucial feature here is that the parity (odd

or even nature) of the permutations must be recognized and only odd permutations are considered. It is this property which leads to many of the seemingly pathological cases that confound many attempts at rigorous description of stereochemistry. Having done this each potential stereocenter with symmetrically related substituents is checked to see if those substituents themselves contain potential stereocenters. If they do not, then the node to which they are attached can never exhibit configurational stereochemistry and is flagged as such. Thus a carbon atom with two methyl substituents would be found not to possess stereochemistry in this way. The procedure of checking potential stereocenters is done iteratively, as long as new nonstereocenters are found. Since multiply-bonded atoms have already been processed to look like tetravalent saturated atoms, they are treated similarly here. The output of this module is the Stereochemical Connection Table (SCT) which includes only those atoms which are capable of exhibiting configurational stereochemistry. Atoms which were rejected as stereocenters by this module are retained for use in reducing the size of the relevant symmetry group as described in the next section. Since the number of potential stereoisomers increases as 2^m where m is the number of potential stereocenters, reducing the size of m to the minimum necessary is a substantial efficiency both in time and storage. (see examples)

2.1.4 Configurational Symmetry Group

The purpose of this module is to determine the effect of the permutations in the symmetry group on the potential stereocenters. This representation of the symmetry group is necessary for the generator to work properly. The basic part of this module is largely unchanged from last year's version as described in the previous annual report. Two modifications have been made since then. The first is that the symmetry group is processed here as elsewhere in the program as two separate pieces for the reasons described above. Second, it was found that a substantial saving could be made by reducing the size of the symmetry group to that subgroup (technically a homomorphic image) which is concerned only with the potential stereocenters. This is done by eliminating those permutations which only effect parts of the molecule which do not exhibit any configurational stereochemistry. Since these parts of the molecule were themselves found earlier by just these permutations, it is a relatively easy matter to discard them afterwards. The resulting symmetry group is reduced by(at least) a factor proportional to 2^r where r is the number of "rejected stereocenters". This leads to a significant savings in time since the symmetry group must be scanned through several times when stereoisomers are generated.

2.1.5 Generator

This module takes the SCT and CSG and generates all the possible stereoisomers. The basic workings of this program are as described in the previous annual report. Modifications were necessary to accommodate the two part symmetry group as described above. Two new features have also been added here. First, the program is capable of detecting enantiomeric pairs of stereoisomers based on the configuration of the stereocenters. This does not include cases where enantiomerism results from conformational or other structural features. Second, the program is capable of computing the symmetry group of each stereoisomer. In general this will be a much smaller group than the CSG for each individual stereoisomer. These two features were added in anticipation of their need later on when capabilities for constrained stereoisomer generation become available. Interpretation of spectral properties such as proton and carbon nmr generally require knowledge of the symmetry group of the stereoisomer being examined. At this stage the outputted stereoisomer is in a canonical form based on the input numbering of the original CT. Because of the very compact representation possible for stereoisomers discussed in last year's annual report, this canonical form is simply an integer from 0 to 2^n where n is the number of stereocenters. Some future plans for the more transparent output required are discussed in the section on future plans. (See example.)

2.1.6 Examples

Several examples are provided here to demonstrate some of the capabilities of the program.

Example 1. The first is 3-6-dimethyl-4-octene, a simple hydrocarbon which exhibits double bond and configuration stereochemistry and has a reduced number of stereoisomers due to symmetry.

3-6-dimethyl-4-octene

```
1 0 0 0 2
2 0 0 1 3
3 0 5 4 2
4 0 3 0 0
5 0 3 11 12
6 11 12 7 0
7 6 8 9 0
8 0 0 0 7
9 0 0 7 10
10 0 0 0 9
11 5 6 0 0
12 5 6 0 0

 THE SCT:

3 0 5 4 2
7 6 8 9 0
5 0 3 11 12
6 11 12 7 0

STEREOCOUNT= 6

 THERE ARE 6 STEREOISOMERS
0 1 1 -1
1 0 1 1
2 0 1 1
4 1 1 -1
5 0 1 1
6 0 1 1
```

Five separate output results are given for this example:

1) The first twelve rows are the Multiple Bond Connection Table (MBCT). The first number is the atom number and the following four are the atoms to which it connects. (0 is hydrogen) Rows 11 and 12 are correspond to the fictional nodes which label the edges of the double bond.

2) Next is shown the Stereochemical Connection Table (SCT). The program has found the two asymmetrically substituted carbons (3 and 7) and the double bond (5 and 6).

3) A counter (discussed below) has determined that there are 6 distinct stereoisomers. This is the STEREOCOUNT.

4) The generator has likewise determined that there are 6 stereoisomers.

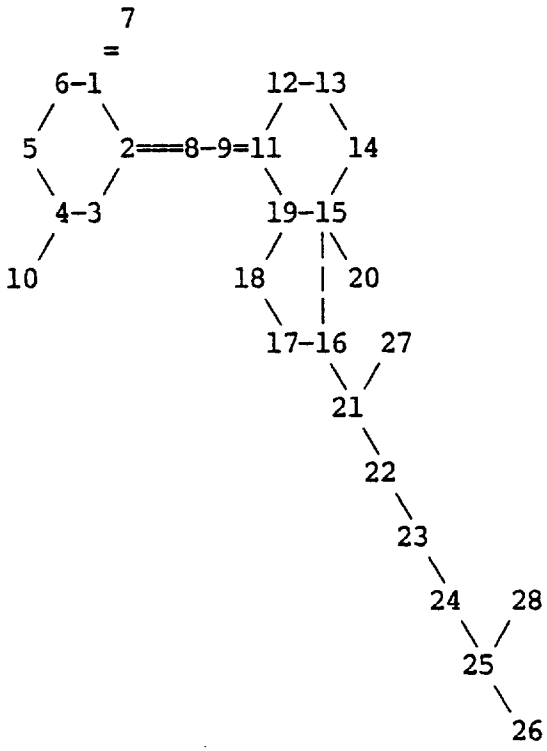
5) The stereoisomers are listed. The first number on each row is the canonical label for each. The correspondence is:

0 R-S-trans
1 S-S-trans
2 R-R-trans
4 R-S-cis
5 S-S-cis
6 R-R-cis

The second number on each row tells whether this particular stereoisomer is achiral (1) or has an enantiomer (0). Enantiomeric pairs are listed on consecutive rows. The final two numbers on each row indicate the symmetry group of each stereoisomer. Those with 1 1 have rotational symmetry and those with 1 -1 have a plane of symmetry.

Example 2. The second example is Vitamin D3 and is included here to illustrate the capabilities of the program in finding stereocenters.

Vitamin D3



Atom number 10 is Oxygen, the rest are Carbon.

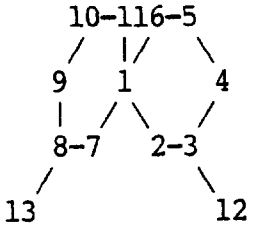
THE SCT:

```
4 10 3 5 0
15 20 19 14 16
16 21 15 17 0
19 15 11 18 0
21 27 16 22 0
2 29 30 1 3
8 29 30 9 0
9 31 32 8 0
11 31 19 32 12
25 28 24 26 0
```

STEREOCOUNT= 128

THERE ARE 128 STEREOISOMERS

For this example only the SCT and number of stereoisomers are shown. The first 5 rows correspond to the 5 asymmetrically substituted carbons. The next four rows correspond to the 4 doubly-bonded atoms which can exist in distinct cis and trans forms. The final row corresponds to the gem-dimethyl substituted carbon on the side chain. This is retained for the reasons discussed above. Both the counter and the generator have established that there are 128 stereoisomers (the theoretical maximum). Example 3. The disubstituted spiro-undecane shown below has only one element of symmetry, the "rotation" axis through carbon 1. This is an even permutation so that carbon 1 remains a stereocenter. NST is the number of stereocenters, NDBAT is the number of doubly-bonded atoms and NRJ is the number of stereocenters rejected by the prefilter.



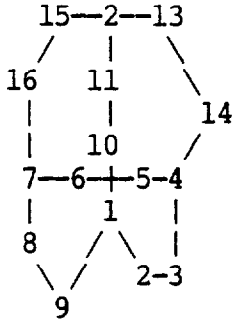
THE SYMMETRY GROUP HAS ORDER P= 2
NST= 3 NDBAT= 0 NRJ= 0

STEREOCOUNT= 6

THERE ARE 6 STEREOISOMERS

```
0 0
7 0
1 0
6 0
2 0
3 0
```

Example 4. The hydrocarbon shown below is the higher homolog of adamantane. The conformational process of turning the structure "inside-out" interconverts the structure with all the hydrogens pointing inside the cage with the structure with all the hydrogens pointing out. The same process interconverts the 3 out 1 in structure with the 1 out 3 in.

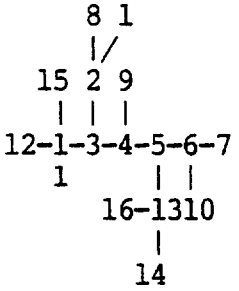


```
THE SYMMETRY GROUP HAS ORDER P= 24
 NST= 4 NDBAT= 0 NRJ= 0

STEREOCOUNT= 3

THERE ARE 3 STEREOISOMERS
 0 1
 1 1
 2 1
```

Example 5. The substituted heptane shown below has two extensively branched symmetrically related substituents at the central carbon. The program detects that this structure can have only 1 stereoisomer and prints this out rather than going through the counting and generating procedures.



```
THE SYMMETRY GROUP HAS ORDER P= 128
 NST= 0 NDBAT= 0 NRJ= 7 THERE IS 1 STEREOISOMER
```

2.1.7 Counter

Another new feature of the program is a procedure which counts the number of stereoisomers for a structure without generating them by using the CSG and the appropriate

combinatorial theorem. This represents the first solution to the problem which dates back to the 1870's. Since the counter works much faster than the generator, this is a very useful feature as the number of stereoisomers can be obtained quickly if only this is needed. This differs from the structure generator where a faster counter was not possible. In addition, having the counter and the generator working independently allows a mutual checking for bugs during development of the program since the two results must be the same for any test case.

2.1.8 Interface to CONGEN

The current interfaced version of the stereogenerator with CONGEN is intended primarily for testing purposes and does not represent the final version. The stereogenerator runs as a separate SAIL fork which is started only when the STEREO command is issued. The desired structure is constructed as a pattern in EDITSTRUC. The CONGEN command: STEREO (name) starts the fork and the stereogenerator. The program asks for an output file and then returns a brief summary of the results to the terminal and a more complete set of results is written on the file. On termination of the generator, control returns to CONGEN.

2.1.9 Future Plans

The following features (at least) will be added to the existing program:

1) Designations of stereocenters as either R or S based on constitutional priorities only. This will be for aid in interpretation only as these designations are not useful internally to the program.

2) Recognition of cis and trans double bonds for the same reason.

3) Stereoisomer output which is interpretable and compatible with character terminal output. This will most likely be done in conjunction with the existing drawing program. The compatibility with character based terminals is a strength of CONGEN at present.

4) Versatility in the handling of the stereochemistry of atoms other than carbon. In particular there should be a choice as to whether a nitrogen atom is thought to be able to invert freely.

The second stage of the development in this effort is to give CONGEN the ability to constrain stereoisomer generation. The algorithm of the generator was designed so that a number of useful constraints, particularly concerning relative

stereochemistry between stereocenters can be applied prospectively. That is, the undesired stereoisomers would not be generated. Other constraints, such as those which involve the symmetry of the stereoisomers can be applied during the generation. Finally, there will certainly be some constraints which have to be applied after generation.

2.2 Constraints Interpretation

The area of automatic interpretation of constraints in CONGEN structure elucidation problems is interesting and important for two reasons: 1) we want to free the chemist as much as possible from having to understand CONGEN's method of building structures; and 2) problems can be solved much more efficiently if CONGEN can perform some preliminary examination of them and find an alternative, efficient way to solve the problem. Our first efforts in this direction have resulted in what we call the "GOODLIST interpreter", which employs the method of constructive substructure search as described in the following sections. The GOODLIST interpreter is designed to make more efficient use of information about required (GOODLIST items plus Superatoms) structural features of an unknown molecule.

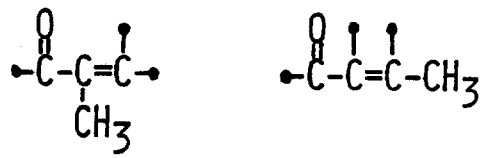
2.2.1 Abstract of Method

We present a solution to the problem of constructing all structural isomers of a given empirical formula given also a set of required partial structures which overlap, i.e., share atoms in common, to an unknown extent. Our method takes a collection of non-overlapping partial structures (in the limit, all atoms in the empirical formula) and, using a technique we term "constructive substructure search," determines the set of subproblems which incorporate all given partial structures, including all possible overlaps, required to be present in each isomer. Each subproblem is solved in turn by CONGEN to yield finally the complete set of isomers, e.g., structural candidates for an unknown compound. Our method allows facile solution of certain structural problems which are beyond the scope of other computer-based methods.

2.2.2 Introduction to Method

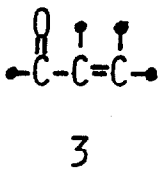
It is characteristic of structure elucidation based on data from physical and chemical methods that much structural information is redundant. Physical methods, for example, are frequently complementary. One technique provides structural information which can be used to elaborate information gathered by another. The collection of partial structures present in an unknown derived by such methods frequently contain atoms or

groups of atoms shared among two or more partial structures. Chemists must take this into account when considering how the partial structures might fit together to yield the structure of an unknown compound. As a simple example, the carbon-carbon double bond of an inferred vinyl methyl functionality may or may not be the same as the double bond of an inferred , -unsaturated ketone. As long as the empirical formula admits of two (or more) double bonds and in the absence of additional information, both possibilities must be considered. Therefore, the chemist will consider 1, 2 and 3,4 as tentative building blocks for further elaboration of the example structure.

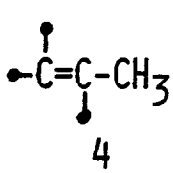


1

2



3

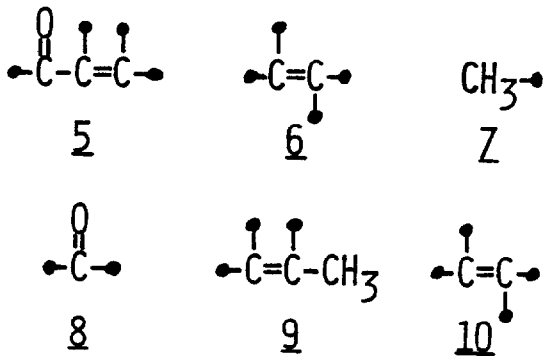


4

Although computer programs, including CONGEN, now exist to assist chemists in constructing structural isomers based on information about partial structures, the programs have one serious limitation in common. Each program must use as building blocks non-overlapping structural fragments. This limitation leads to at least two important problems; 1) The chemist using such a program must select non-overlapping partial structures; otherwise an incomplete set of structures will result. This manual procedure is time-consuming, unnatural and prone to error; and 2) as a consequence of (1), problems are solved less efficiently by the program because the detailed environment of fewer atoms is specified to ensure the absence of overlaps. Thus, undesired structures are built only to be discarded upon later evaluation. We feel that a solution to the first problem is extremely important. Our experience is that there are already sufficient barriers to use of computers as assistants in problem-solving. We feel strongly that allowing a chemist to input structural information freely without regard to overlapping partial structures would reduce that barrier. The importance of the second problem is that certain structural problems become difficult or impossible to solve with current programs (that is,

impossible in the sense that resources of computation, time and money are finite).

For the example cited above, current programs would be forced to consider for completeness a starting point of either 5,6,7 or 8,9,10.



Assuming that the problem involves other partial structures or atoms, either starting point results in construction of structures including 1, 2 and 3,4 together will many other structures which do not obey the constraints on the problem. Application of constraints in CONGEN is automatic, but the retrospective testing of every structure for desired structural features which could not be used to begin with is very inefficient.

We sought, therefore, a method which would emulate the manual approach to the problem of determining structural candidates based on overlapping partial structures. Stated in the simplest terms, the method should translate the constraints on desired structural features, or GOODLIST constraints, into new sets of partial structures which incorporate the features at the beginning of the structure generation procedure. Such a method would translate automatically the constraints in the problem mentioned above to yield three new problems represented by 1, 2 and 3,4. Subsequent sections describe a method which performs this translation. We illustrate the method with examples drawn from our own work, some of which could not be solved in reasonable time using existing programs.

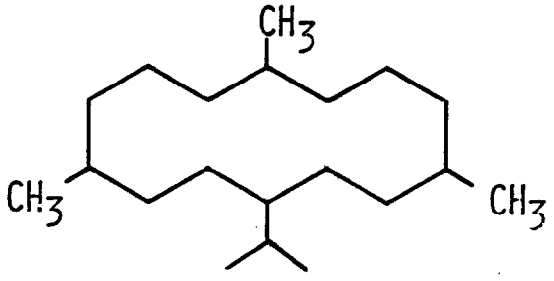
2.2.3 METHOD

There are usually many constraints on a structural problem brought to CONGEN, including those implied by other constraints. Manual approaches to structure elucidation involve recognition of implied constraints and resolution of overlapping partial structures (mentioned above) as structural candidates are constructed. The translation of constraints to discern their implications and elaboration of those implications into more efficient statements of a problem involves complex reasoning about chemical structures. This reasoning is susceptible to analysis and encoding in a computer program.

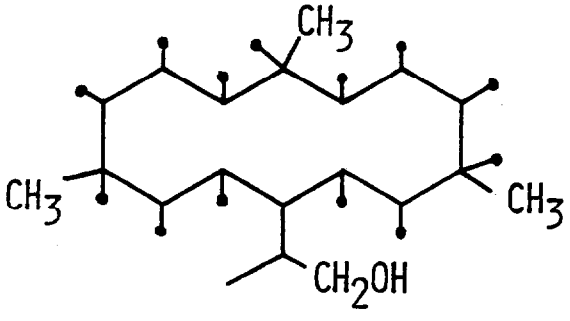
Our initial experiment in constraints interpretation involved determination of the implications of designated numbers of hydrogens associated with particular atoms. Translation of this information reduces many problems to triviality, for example, "construct all isomers of $C_{20}H_{44}N_2$ which possess no methyl groups". We describe below the next step in our efforts, a method for translation of desired, or GOODLIST, structural features.

Our method is based strongly on our observations of how chemists actually solve the problem of using overlapping partial structures. We introduce the method with an example which in fact provided the basis for the first programming efforts.

The structural problem involved an unknown compound of empirical formula $C_{20}H_{34}O_1$. The compound was isolated together with other cembranolides, therefore the assumption was that the unknown possessed the unrearranged cembrane skeleton (11).



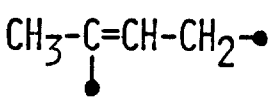
11



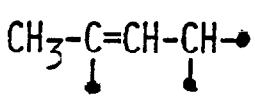
12

These data indicate that the structure is based on the

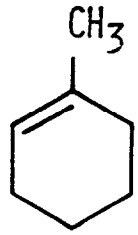
skeleton 12 together with allocation of three new bonds in such a way as to yield the desired partial structures 13-15. (Bonds with an unspecified terminus, or "free valences" in 12 may be to any atom including hydrogen, while in 13-14 the indicated free valences are specified to be to non-hydrogen atoms.)



13



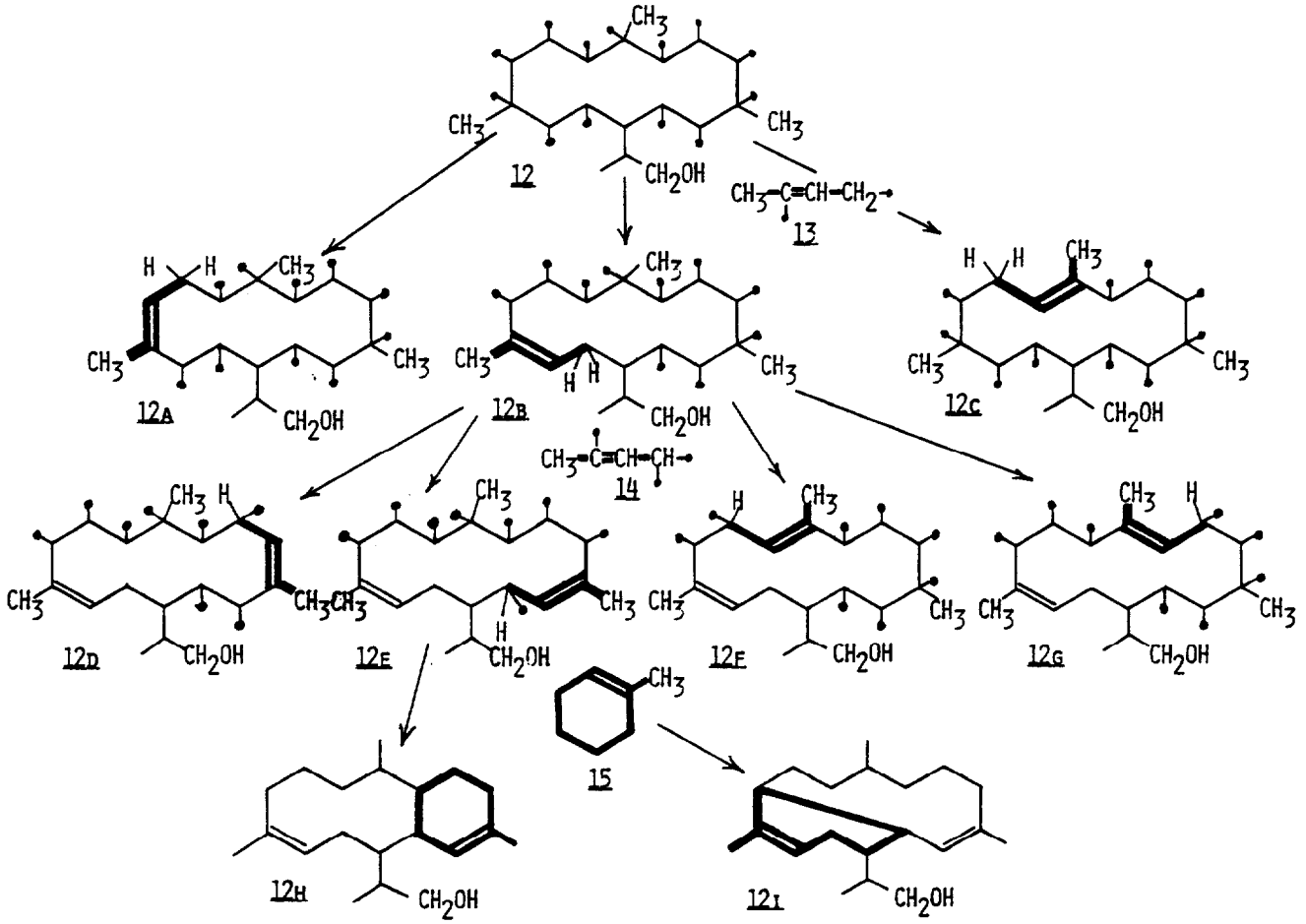
14



15

In this problem, the skeleton, 12, possesses all non-hydrogen atoms of the empirical formula. Thus, the substructures 13-15 overlap completely with 12 (and partly with each other). A conventional approach to this problem would allocate three new bonds to 12 in all possible ways and test each result against the GOODLIST constraints 13-15. There are many thousands of possible allocations and the computational task of building and testing each one was so time consuming it was terminated. The chemist then retired to his desk and, using pencil and paper, in a short time determined the seven possible structures obeying the constraints.

It is clear conceptually how such problems are solved. It is obvious, considering the topological symmetry of 12, that there are only three places in 12 where 13, for example, might fit, or match. The three matchings 12a-12c are shown below. Each matching consumes two free valences to form the new double bond and effectively places a hydrogen on the terminal atom of the substructure yielding the required $-CH_2-$ group. For each matching of 13, there are several ways to fit in the next GOODLIST substructure, 14. There are four ways to perform this matching for 12b, resulting in 12d-12g, below. Again, a pair of free valences is consumed to construct the new double bond. In this case, however, the substructure 14 terminates in a methine group, effectively leaving a bonding site open (see 12d-12g) which must be used in forming a new bond in a subsequent step. Incorporation of the final GOODLIST constraints, 15, proceeds by creation of a new bond (with the methine, above, as one terminus) to yield a six-membered ring possessing a double bond. Certain structures, e.g., 12f, yield no results because a bond cannot be formed which meets the requirements of 15, while 12g yields two results 12h-12i, as shown below.



In this example, some matchings result in construction of new bonds to form the extra double bonds and ring of the unknown. In the general case, the procedure is constructive in that bonds are formed to new atoms or substructures to obtain partial structures which are required. Using the method described below in conjunction with CONGEN, we can determine automatically and quickly the seven solutions.

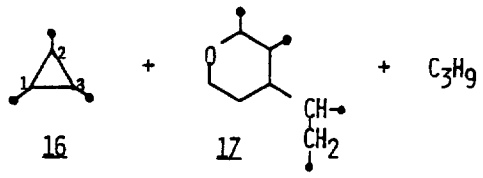
2.2.4 GOODLIST Constraint Interpretation Search

Our method emulates the manual method by searching for ways to map possibly overlapping GOODLIST substructures into the partial structures and/or atoms in the initial problem formulation. The method, illustrated schematically below, includes the following steps.

2.2.4.1 Formulation of the Initial CONGEN Problem

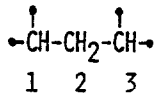
The initial structural problem is defined to be a set of non-overlapping partial structures, or "Superatoms," plus the remaining atoms in an empirical formula (below). Thus, specifications of the initial problem can proceed just as with current use of the program. However, a wide variety of initial specifications is possible, from initial problems where all atoms are part of a superatom (e.g., 12, above) to the limit of simply the empirical formula (where all atoms are of course non-overlapping). For example, the problem of the cembrenolide outlined above is solved with little difference in efficiency beginning with the empirical formula and utilizing 13-15 as GOODLIST constraints. In the example below, assume that partial structures 16 and 17 are known to be non-overlapping superatoms, leaving C_3H_9 remaining from an empirical formula $C_{13}H_{22}O_1$.

INITIAL CONGEN PROBLEM



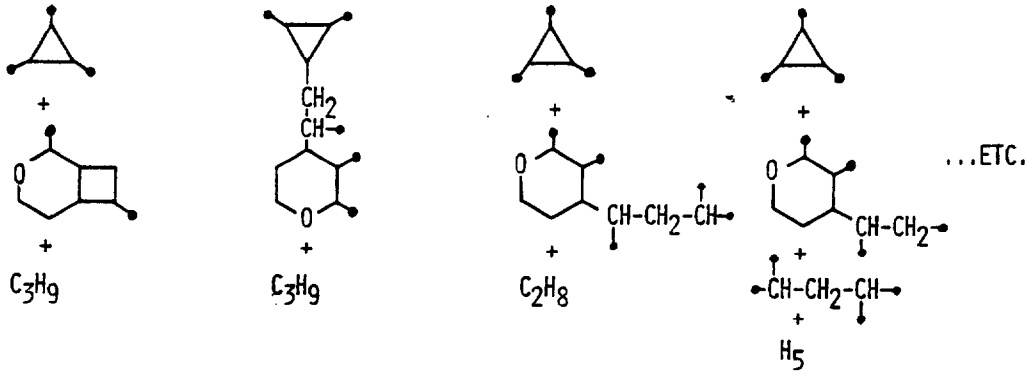
16 17

GOODLIST CONSTRAINT



18

NEW CONGEN PROBLEMS



A B C D

2.2.4.2 Constructive Substructure Search

Assume that substructure 18 is known to be present in a molecule of unknown structure with no additional information on possible overlaps with 16 and 17. The method begins by finding

all ways in which the GOODLIST substructure (18) can be constructed using Superatoms and atoms in the initial problem.

There may be several ways to incorporate a given GOODLIST constraint in a CONGEN problem. The substructure may be incorporated by forming bonds within a substructure (yielding A), forming new bonds between (or among) substructures (yielding B), forming bonds between substructure(s) and remaining atoms (yielding C) or construction of the substructure wholly from remaining atoms (yielding D).

The result of constructive incorporation of each GOODLIST substructure is a set of new CONGEN problems. Our stepwise procedure continues by incorporating the next GOODLIST item in a depth-first generation scheme. For example, considering the cembrenolide, above, one of the three new problems after incorporation of 13 is chosen for the next step, incorporation of 14. One of the resulting problems is chosen for incorporation of 15. The procedure continues until all GOODLIST items have been incorporated or until the next GOODLIST item cannot be built from superatoms and atoms in the current problem. In the latter case, the program backtracks one step and tries the next problem at the previous level.

2.2.4.3 Obtaining Final Structures

The results of the constructive procedure may be complete structures, for example, 12h and 12i. Usually, however, the result is a set of incomplete problems. Each problem includes superatoms and remaining atoms which are guaranteed to be non-overlapping and which contain all desired structural features. The standard CONGEN procedure for structure generation can then be invoked. However, the task of testing for substructure and ring constraints is simplified in that GOODLIST constraints are already incorporated.

2.2.5 Limitations

There are some limitations to the procedure which decrease its efficiency compared to what might be possible with further work. One limitation is the problem of duplication inherent in the procedure. Although many steps are taken to perceive and utilize topological symmetry in the constructive substructure search, there remains the possibility of constructing duplicate CONGEN problems whenever the constructive procedure creates symmetries which were not present originally. Therefore, we convert each CONGEN problem to a canonical form and compare problems to eliminate duplicates. Another potential source of duplication is construction of duplicate (isomorphic) final structures from different CONGEN problems. Again, canonicalization serves to prevent presentation of duplicate structures to the chemist.

A second limitation is related to the absence of a mechanism for preventing the association of atoms in a GOODLIST substructure with atoms in a CONGEN problem. It may be known that a GOODLIST substructure does not share atoms (i.e., overlap) with one or more superatoms (i.e., some spectroscopic evidence is available to distinguish them). However, there is no mechanism for preventing association of atoms in a superatom with atoms in a GOODLIST item. Some undesired structures result which must be removed by subsequent tests.

2.2.6 Future Directions

The program described in this section will be incorporated in the existing CONGEN program in such a way that it will be invisible to the chemist using the program. Initially, the GOODLIST substructures specified as constraints will be incorporated automatically at the beginning of the problem as described above. Within a short time, the method of specification of a problem will be changed to include only the empirical formula together with inferred partial structures without regard to overlaps, leaving to the program the task of determining those overlaps and specifying the set of problems to solve.

Automatic interpretation of GOODLIST constraints is only the first phase of our efforts. Incorporation of BADLIST (undesired structural features) substructures in the procedure is a necessary next step. Subsequently we will attack the problem of discerning constraints which are implied by the input data, including detection of unclear or ambiguous statements about a structure. The constraints interpreter should be capable of a dialog with the chemist using CONGEN to clarify such points prior to structure generation.

2.3 Experiment Planning Program

Now that Congen gives us the capability of constructing all plausible candidates under an initial set of constraints, the next problem is to provide the chemist with some assistance in rejecting incorrect candidates and focussing on the correct structure. This process must involve the examination of the candidates to determine their common and unique features, and the designing of experiments to differentiate among them.

The initial work on this problem has begun by providing a new function, the EXAMINE function, which gives a chemist the ability to survey sets of structures for particular combinations of substructures, ring-systems etc. This function has now been incorporated into the CONGEN program; details and examples are given later.